一种高效安全的密文电子病历多关键字检索方案

邓 兰 徐 进

（华中科技大学同济医学院附属同济医院 武汉 430030）

〔摘要〕 提出针对密文电子病历的关键字检索方案，介绍相关基础知识，阐述系统模型、方案算法，测试分析该方案安全性及可行性，指出该方案可在保持电子病历安全性与隐私性的同时提升实用性，使云环境下电子病历的密文存储成为可能。

〔关键词〕 电子病历；安全；可搜索加密；云存储

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.01.009

An Efficient, Secure and Multi－keyword Retrieval Scheme for Encrypted Electronic Medical Records

DENG Lan, XU Jin, Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology, Wuhan 430030, China

〔**Abstract**〕 The paper proposes a keyword retrieval scheme for encrypted Electronic Medical Records (EMR), introduces the relevant basic knowledge, expounds the system model and scheme algorithm, tests and analyzes the security and feasibility of the scheme, and points out that the scheme can improve the practicality while maintaining the security and privacy of EMR, and make it possible to store encrypted EMR in the cloud environment.

〔**Keywords**〕 Electronic Medical Records (EMR); security; searchable encryption; cloud storage

1 引言

随着医疗信息化建设和医药卫生体制改革的推进，电子病历得到广泛应用。云计算技术的日益成熟和医疗数据互联互通需求增强，促使电子病历存储在云服务器成为医疗信息化未来发展趋势。云存储应用既方便多方机构存储和交换数据，又能大幅节省医院构建和维护独立数据库的开支。电子病历作为医疗卫生信息的主要载体包含大量有价值信息，如患者身份信息、联络方式、治疗信息等，通过网络存储和传输时存在隐私泄露、非法访问等隐患[1]，一旦电子病历信息被泄露可能会给患者和医院带来较大损失[2]，影响医患关系。目前对电子病历安全性和隐私性的保护措施主要包括身份认证[3]、操作日志、数据脱敏[4]、数据加密[5-6]等。其中数据加密是保护电子病历安全性的重要手段，一经加密即使黑客非法窃取也难以解密获取信息。然而数据保密性与可用性之间难以平衡[7]是电子病历加密存储难以推广的重要原因，主要表现在电子病历加密后难以对其

〔修回日期〕 2021－06－12

〔作者简介〕 邓兰，硕士，工程师，发表论文5 篇；通讯作者：徐进，硕士，工程师。

〔基金项目〕 国家重点研发计划“规范化、全周期重大出生缺陷大数据平台建设”（项目编号：2019YFC1005100）。

进行检索。电子病历数据一经加密，检索时需要对整个数据库解密，计算和通信成本过高。本文提出一种对密文电子病历进行关键字检索的方案，利用布隆过滤器和 k 最邻近（k - Nearest Neighbor，kNN）加密等工具，将可搜索加密中的正向和反向索引相结合，构造一个两级索引，使用户能够对密文电子病历进行多关键字检索且检索效率较高，从而平衡电子病历的保密性与可用性，使电子病历密文存储成为可能。

2 基础知识

2.1 可搜索加密

2.1.1 概述 2000 年 Song D X、Wagner D 和 Perrig A 提出可搜索加密问题[8]，2004 年 Dan B、Crescenzo G D 和 Ostrovsky R 等[9]提出可关键字检索的公钥加密（Public - Key Encryption with Keyword Search，PEKS）概念并提出第 1 个 PEKS 方案。可搜索加密最早应用于加密电子邮件的关键字检索，随着云计算的兴起，其在云存储领域得到广泛研究[10-12]。可搜索加密使用户能够在不全部下载并解密数据且不让云服务器获取数据和关键字明文的前提下对数据进行关键字检索。

2.1.2 基于索引的方案 较常用的可搜索加密方案[13]，即通过为数据额外生成 1 个安全索引，索引中记录关键字与数据的对应关系，从而实现关键字检索功能。索引分为正向和反向索引。正向索引是由数据条目指向关键字，优点是便于更新，缺点是在检索某个关键字时需要扫描整个索引，检索效率较低。反向索引则为每个关键字存储 1 个数据 id 列表，检索时可以直接获取某个关键字的检索结果，效率较高，但索引更新较麻烦且不便于支持多关键字检索。

2.1.3 多关键字检索 即搜索同时包含多个关键字的数据，考虑到数据量庞大，支持多关键字检索是提高检索效率、准确性和使用友好度的必要功能。

2.2 布隆过滤器

2.2.1 概述 布隆过滤器由 Bloom B H[14]于 1970 年提出并由 Eu - jin Goh[15]首先应用于可搜索加密中。布隆过滤器使用 1 个 0、1 数组代表集合 S，S 中的每个元素都通过 k 个哈希函数映射数组中的 k 个位置，这些位置被设为 1，其他位置则为 0。当检测 1 个元素是否属于 S 时只需计算出该元素对应的 k 个位置并查看这些位置是否全为 1 即可，见图 1。

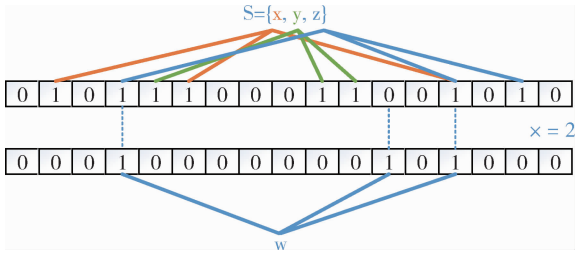

图 1 布隆过滤器原理

2.2.2 原理 集合 S 中包含 x、y、z 3 个元素，每个元素通过 3 个哈希函数对应 3 个位置，这些位置全部被置为 1，可得出代表集合 S 的布隆过滤器。为检验元素 w 是否存在于集合 S 中时，通过计算 w 的 3 个哈希值，也得到 1 个布隆过滤器。两个布隆过滤器的内积为 2，小于 $k=3$，可知 w 并不属于集合 S。布隆过滤器具有查询空间和时间效率较高的优点，因此可将其应用于密文电子病历检索，为每条电子病历生成 1 个布隆过滤器，在实现电子病历密文存储的同时保留关键字检索功能。

2.3 kNN 加密

布隆过滤器虽然能够实现搜索功能，但其如果以明文形式存储是不安全的，攻击者可以轻易获知 1 个布隆过滤器包含哪些关键字，因此需要将其加密。kNN 加密算法由 Wong W K、Cheung W L 和 Kao B 等[16]提出，用于在加密情况下计算两个数据项的相似性。kNN 加密使用 1 个比特向量 K 和两个可逆矩阵（M_1，M_2）作为密钥。对于两个布隆过滤器 I 和 Q，首先使用 K 和随机数 r 将 I 分成两个随机向量（I_1，I_2），再用（M_1，M_2）将其加密得到（$M_1^T \cdot I_1$，$M_2^T \cdot I_2$），即加密的布隆过滤器 I。对于与 I 比对的 Q，使用 K 和另一个随机数 r'将其分成（Q_1，Q_2），再用（M_1，M_2）将其加密得到（$M_1^{-1} \cdot Q_1$，$M_2^{-1} \cdot Q_2$）。最后进行如下运算可得出 I 和 Q 的内积[9]：

$$M_1^T \cdot I_1 \cdot M_1^{-1} \cdot Q_1 + M_2^T \cdot I_2 \cdot M_2^{-1} \cdot Q_2 = I_1^T \cdot Q_1 + I_2^T \cdot Q_2 = I \cdot Q$$

3 方案

3.1 系统模型

本系统成员包括电子病历管理员、使用者以及云服务器。管理员将电子病历加密并生成安全索引存储到云服务器中。使用者检索时只需发送关键字给管理员，由管理员使用密钥生成该关键字对应的搜索凭证并回复给使用者，其中管理员和使用者之间的身份认证不在讨论范围。云服务器收到搜索凭证后将其与安全索引做一定计算即可得出搜索结果并返回给使用者，见图 2。在本方案中电子病历加密和索引生成是分开的，电子病历使用常规的对称加密算法，如 AES、SM1、SM4、SM7、祖冲之密码等。

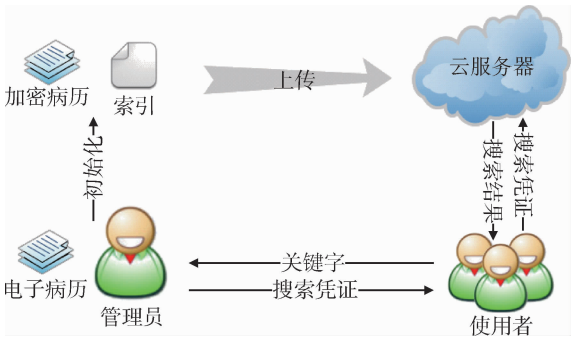

图 2 系统模型

3.2 方案算法

3.2.1 *初始化* 管理员生成方案所需密钥包括伪随机函数 $F(\,)$ 的密钥 K_1、K_2，以及 kNN 密钥 $K_{kNN}=(K, M_1, M_2)$。对于每条电子病历提取出若干关键字。对于每个关键字 w，使用伪随机函数 $F(\,)$ 和密钥 K_1 与其进行计算，得到 tag_w。使用 $F(\,)$ 和 K_2 生成 K_w，用 K_w 将 w 对应的每个病历 id 加密成 $E(id)$，与 tag_w 构成 1 个反向索引。对于每条电子病历，用其包含的所有关键字生成 1 个布隆过滤器。对每个布隆过滤器使用 kNN 加密算法和密钥 K_{kNN} 将其加密，所有加密的布隆过滤器即构成正向索引。将索引 1 和 2 连同加密的电子病历存储到云服务器中，见图 3。

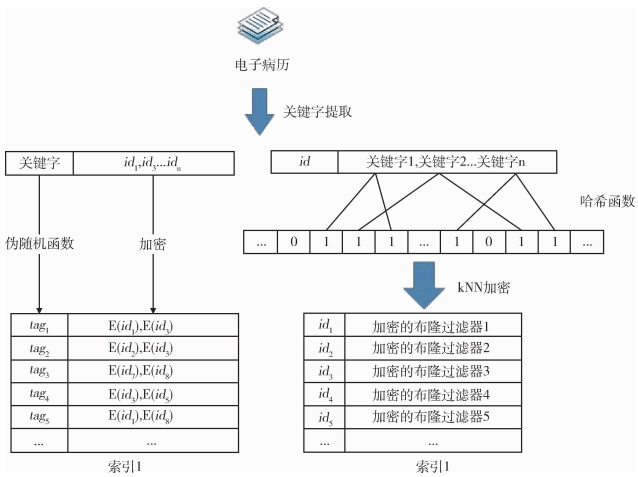

图 3 初始化流程

3.2.2 *生成检索凭证*（图 4）

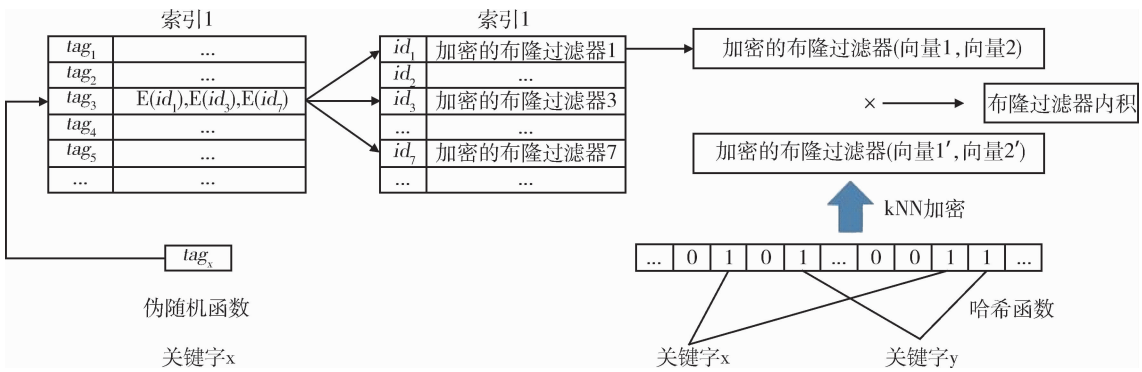

图 4 检索流程

从需要搜索的多个关键字中挑选出 1 个，如 x，使用伪随机函数 $F(\,)$ 和密钥 K_1 对其进行处理得到 tag_x。使用 $F(\,)$ 和 K_2，生成关键字 x 的 id 加密密钥 K_x。使用哈希函数将需检索的所有关键字映射到 1

个布隆过滤器中。使用 kNN 加密算法和密钥 K_{knn} 对该布隆过滤器加密。将生成的搜索凭证即 tag_x、K_x 和加密的布隆过滤器发送给云服务器。

3.2.3 密文检索 云服务器先根据 tag_x，在索引 1 中找到关键字 x 对应的所有病历 id，使用 K_x 将其解密。对于每个病历 id，从索引 2 中获取其对应的加密布隆过滤器，将搜索凭证中的加密布隆过滤器与每个 id 对应的加密布隆过滤器进行计算，得到的内积即是要搜索的关键字与每条电子病历的相似度。将内积值大于某个阈值的密文电子病历返回给使用者。

4 实践与分析

4.1 概述

在真实病历数据基础上用实验测试方案，方案可行性、安全性得到验证。实验证明该方案能够实现对密文电子病历的多关键字检索功能，且检索准确性和检索效率较高。

4.2 安全性

由于信息均被加密存储，攻击者无法从索引中获取信息。此外进行检索时由于 kNN 加密中使用 1 个随机数，每次生成搜索凭证都不相同，攻击者无法判断两次搜索的关键字是否完全相同。通过模拟攻击发现无法获取病历明文、关键字明文、未被检索的病历 id、两次检索的关键字是否相同等关键信息，安全性得到保障。

4.3 时间和空间效率

4.3.1 概述 以华中科技大学同济医学院附属同济医院 3 个月的病历数据为例。该院 2020 年 12 月住院电子病历数为 24 086 份，11－12 月为 44 267 份，10－12 月为 61 200 份。每份病历使用“住院号－住院次数”作为 id 并提取性别、年龄、籍贯、主诉、既往史、诊断等 50 个关键字，按照公式 $m = -\frac{n\ln p}{(\ln 2)^2}$，$k = \frac{m}{n}\ln 2$（$p$ 是误判率，n 是关键字个数，m 是布隆过滤器长度，k 是哈希函数个数），布隆过滤器长度取 500，哈希函数个数取 10。然后通过程序为上述病历数据构建索引并进行检索操作，输入关键字后程序将返回包含这些关键字的病历 id，方案时间、空间成本，见表 1。

表 1 方案时间、空间成本

项目	1 个月	2 个月	3 个月
病历数量（份）	24 086	44 267	61 200
关键字数量	66 137	107 959	139 771
索引 1 生成时间（秒）	1 396	2 519	3 579
索引 2 生成时间（秒）	1 224	2 253	3 160
索引生成总时间（秒）	2 620	4 772	6 739
检索凭证生成时间（秒）	1	1	1
检索时间（秒）	1	1	2
检索凭证大小（KB）	4.2	4.2	4.2
索引大小（KB）	110 361	202 438	281 103

4.3.2 时间成本 耗时最长的步骤是索引生成，耗时与病历数量成正比。由于索引只需生成 1 次且索引 2 的生成可分散到每份电子病历提交时进行，因此是可接受的，更高性能的计算机可缩短生成时间。生成检索凭证耗时恒定，基本可以忽略。关键字检索耗时与被检索第 1 个关键字对应的病历数相关，与病历总数无关，第 1 个关键字对应的病历数越少检索越快。根据方案时间、空间成本统计，检索凭证的生成和检索操作耗时较短，因此检索效率得以保障。

4.3.3 通信成本 进行检索时使用者和云服务器之间只需 1 轮通信且搜索凭证大小恒定，通信成本较低。

4.3.4 存储成本 索引 1 的大小与关键字总数成正比，索引 2 的大小与病历总数成正比，且由于加密后的布隆过滤器由大量浮点数组成，因此需要的存储空间略多，但在云环境下这些存储量是可接受的。此外还可通过适当减小布隆过滤器的长度来缩减索引大小。

5 结语

针对云环境下密文电子病历无法进行关键字检

索的问题，基于布隆过滤器和 kNN 加密算法，结合正向和反向索引，提出一种安全高效且支持多关键字搜索的电子病历加密方案，使电子病历加密并存储在云服务器、实现数据共享成为可能。此外检索方案和电子病历加密方案相互独立，电子病历可使用任何现有方案进行加解密，具有可部署性。目前该方案尚存在存储成本有待降低等不足，需在未来研究中继续完善。

参考文献

1 李晓蓉，宋子夜，任婧怡，等．云计算中基于属性的可搜索加密电子病历系统［J］．计算机科学，2017（S2）：352－357.

2 杨吉江，许有志，王青，等．面向医疗信息的数据隐私保护技术［J］．中国数字医学，2010，5（8）：50－54.

3 张丽，吴海军．基于 PKI 技术的电子病历系统安全研究［J］．科技视界，2020（3）：239－240.

4 臧昊，赵强，卞水荣．基于 XML 的电子病历隐私数据脱敏技术的研究与设计［J］．信息技术与信息化，2017（3）：111－114.

5 高诗，徐薇，王美芹．加密云存储电子病历系统应用研究［J］．中国数字医学，2020，15（2）：15－17.

6 周天舒，朱兴骅，李劲松．面向卫生信息交换的电子病历安全体系设计［J］．中国数字医学，2015，10（6）：83－86.

7 朱勤，骆轶姝，乐嘉锦．数据库加密与密文数据查询技术综述［J］．东华大学学报（自然科学版），2007，33（4）：543－548.

8 Song D X，Wagner D，Perrig A. Practical Techniques for Searches on Encrypted Data［C］. Berkeley：IEEE Symposium on Security & Privacy，2002.

9 Dan B，Crescenzo G D，Ostrovsky R，et al. Public Key Encryption with Keyword Search［C］. Berlin：International Conference on the Theory and Applications of Cryptographic Techniques，2004.

10 Wang B，Yu S，Lou W，et al. Privacy－preserving Multi－keyword Fuzzy Search over Encrypted Data in the Cloud［C］. Toronto：INFOCOM，2014 Proceedings IEEE，2014.

11 Yang Y，Ma M. Conjunctive Keyword Search With Designated Tester and Timing Enabled Proxy Re－encryption Function for E－Health Clouds［J］. IEEE Transactions on Information Forensics & Security，2017，11（4）：746－759.

12 Chen J，He K，Deng L，et al. EliMFS：Achieving Efficient，Leakage－resilient，and Multi－keyword Fuzzy Search on Encrypted Cloud Data［J］. IEEE Transactions on Services Computing，2017（99）：1.

13 李经纬，贾春福，刘哲理，等．可搜索加密技术研究综述［J］．软件学报，2015（1）：109－128.

14 Bloom B H. Space/Time Trade－offs in Hash Coding with Allowable Errors［J］. Communications of the ACM，1970，13（7）：422－426.

15 Eu－jin Goh. Secure Indexes［EB/OL］.［2004－04－30］. https：//www.researchgate.net/publication/2889193_Secure_Indexes.

16 Wong W K，Cheung W L，Kao B，et al. Secure kNN Computation on Encrypted Databases［C］. Providence：ACM SIGMOD International Conference on Management of Data，2009.